DOCUMENT RESUME

ED 269 463 TM 860 317

AUTHOR Schratz, Mary K.
TITLE Measuring Up in an Individualized Way with CAT-ASVAB: Considerations in the Development of Adaptive Item Pools.
PUB DATE Apr 86
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Adaptive Testing; Adults; *Aptitude Tests; *Computer Assisted Testing; *Item Banks; Latent Trait Theory; Microcomputers; Military Personnel; Occupational Tests; Postsecondary Education; *Test Construction; Test Items; *Vocational Aptitude
IDENTIFIERS *Armed Services Vocational Aptitude Battery; Test Specifications

ABSTRACT

A computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) is being developed to test ten areas of vocational aptitude. Development of the CAT-ASVAB is based upon the use of item banks, testing and scoring procedures, assessment of test reliability and test validity, a technical manual, and the development of equated scores for the paper-and-pencil and the CAT-ASVAB. There are three main considerations in the development of item pools: (1) content specifications--careful analysis of the content, skills, and knowledge comprising the aptitude area, and review of items; (2) item preparation, including consideration of the capabilities of the computer and its display, and psychometric standards as judged by experts; and (3) item calibration. The data collected for more than 2,000 test items on a military applicant sample provided item responses for over 2,000 examinees per item, for calibration purposes. The CAT-ASVAB battery uses the three parameter model, which assumes the concept of unidimensionality of each item. Future plans include the use of microcomputers for test administration, further work on the comparability of paper-and-pencil and computerized batteries, and the creation of a central testing facility. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269463

MEASURING UP IN AN INDIVIDUALIZED WAY WITH CAT-ASVAB

Considerations in the Development of Adaptive Item Pools

Mary K. Schratz, Ph.D.

Navy Personnel Research and Development Center
San Diego, California 92152-6800

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

M. K Schratz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 317

Paper presented at the Annual Meeting of the American Educational Research Association, in a symposium on "Measuring Up: The Role of Military Testing," April 1986.

Background

The Department-of-Defense is developing a computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). CAT-ASVAB is intended to replace the paper-and-pencil ASVAB battery, administered to all applicants for military service for selection and classification purposes. The present ASVAB battery is administered to applicants at 68 Military Entrance Processing Stations nationwide and at approximately 1,000 mobile test sites. Aptitude test scores and service-specific composite scores are used to identify candidates eligible for service and to determine their eligibility for particular occupational training programs.

The present ASVAB battery consists of tests in ten aptitude areas. They are: Paragraph Comprehension, Word Knowledge, Mathematics Knowledge, Arithmetic Reasoning, General Science, Mechanical Comprehension, Electronics Information, Automotive and Shop Information, Clerical Speed, and Numerical Operations. The tests are administered in a traditional paper-and-pencil format, i.e. all examinees taking a given test form are administered the same test items. While the aptitude tests included in CAT-ASVAB will, for the present, be the same as for the paper-and-pencil ASVAB, test items will be chosen for administration to an examinee in an individualized way for all power tests comprising the battery. The computerized adaptive testing process is briefly described below.

Computerized Adaptive Testing

The increasing sophistication and availability of large computers has encouraged research on new psychometric procedures for use in the development of tests and in the analysis of test data. An example of one such procedure is "item response theory" (Lord, 1980). Advances in microcomputer technology have provided for increasing sophistication, availability, and portability in design, making computerized test administration and scoring practical for many settings. These two areas of development, item response theory and microcomputer technology, support the computerized adaptive testing process and have encouraged its application.

In adaptive testing, the selection of test items for administration to an examinee is based on his/her performance on previously administered items. The adaptive testing process involves: (1) administering a test item and obtaining the examinee's response; (2) scoring the test item; (3) estimating the examinee's ability level; and (4) selecting a new test item for administration. For each test, then, estimates of measured ability for examinees are obtained via the administration of different test items. Item response theory provides the theoretical framework for the adaptive testing process. The combination of the adaptive testing process, item response theory methods, and computerized test administration and scoring is known as computerized adaptive testing (Weiss & Kingsbury, 1984). This methodology will be implemented for the CAT-ASVAB battery.

Technical Components of CAT-ASVAB

The development of a computerized adaptive testing version of ASVAB involves the formulation of several psychometric procedures and measurement products. Among these procedures and products are: (1) test item banks to support the adaptive testing process; (2) test administration and scoring

procedures; (3) assessment of test reliability and validity; (4) the preparation of a technical manual; and (5) the development of comparable or "equated" scores for the computerized and the paper-and-pencil versions of the battery. Development of all of these psychometric components of the CAT-ASVAB battery is underway; in addition, the design and field testing of a CAT-ASVAB microcomputer delivery system is in progress and scheduled to be introduced later this year. For the purposes of the present paper, considerations in the development of test item pools to support the computerized adaptive administration of this large scale aptitude testing program will be discussed in some detail. Ongoing work on the remaining psychometric and system development components of the CAT-ASVAB battery will be briefly described and future directions will be highlighted.

Considerations in the Development of Adaptive Item Pools

Prominent measurement experts (Green, Bock, Humphreys, Linn, and Reckase, 1984) have indicated that the "main ingredient" for a good CAT is a large, well-distributed item pool with well-estimated item parameters, for a domain with one dominant dimension. For the CAT-ASVAB battery, this "main ingredient" must be developed for each aptitude area measured adaptively. As presently planned, the CAT-ASVAB battery will consist of approximately two hundred test items in each of the nine aptitude areas to be administered adaptively. Items will be assigned to one of two alternate test forms. Producing the "main ingredient" for CAT-ASVAB involves careful consideration of technical issues at all stages of item pool development.

1. Content Specifications

The first step in the development of any item pool, be it for a traditional test or for an adaptive test, is careful analysis of the content, skills, and knowledge comprising the aptitude area. This process is typically accomplished in the development of test content specifications for the battery. In adaptive testing, examinees are administered only those test items suited to their ability level. Therefore, it is essential that the item pool for each aptitude area contain more items than a traditional paper-and-pencil test form, and that the items span a wider range of difficulty. To properly support the adaptive testing process at all ability levels, these considerations must be taken into account in planning the test specifications.

Detailed content specifications have been prepared for the CAT-ASVAB test battery and a comparison of the content represented in both the CAT-ASVAB battery and the paper-and-pencil ASVAB batteries has been made. In developing the computerized adaptive version of ASVAB, domain specifications for each paper-and-pencil aptitude test were broadened to cover the range of examinee ability to be measured adaptively. Textbooks were surveyed from junior and senior high schools, and vocational technical schools to specify the content of each aptitude test. Nearly 4,000 test items were written to cover the measurement objectives represented by the content specifications and these items were pretested on a sample of military recruits. After pretesting, more than 2,000 items were retained for calibration purposes and administered in paper-and-pencil test booklets to military applicants (Prestwood, Vale, Massey, and & Welsh, 1985).

Expert judges were involved in screening the CAT-ASVAB test items during various stages of their development. Their involvement in recent stages of

item pool development, prior to field testing of test items administered via the proposed microcomputer delivery system, has been extremely useful. A panel of twelve expert judges was assembled to verify content designations for individual test items and to perform a comparison, at the item level and at the test level, of potential CAT-ASVAB items and those items presently comprising a paper-and-pencil ASVAB reference form. This panel consisted of test developers and educators with special expertise in the content areas comprising the battery. Documentation of this review/screening process, and the impact it has had on the development of initial forms of the CAT-ASVAB battery and on the development of standards for future CAT-ASVAB forms, will be described in a preliminary CAT-ASVAB Technical Manual scheduled to be completed within the next year.

2. Item Preparation

In preparing test items for inclusion in an adaptive test, several factors must be taken into account. Some of those to be considered are: (1) physical standards, driven by test development design but also by the capabilities of the microcomputer, and (2) measurement standards, implemented through the application of standard item writing principles. With respect to physical standards, consideration of the size and types of displays required for potential items is important to address in developing item pools for adaptive tests. While measurement objectives must always guide the development of items of a particular format, size, or complexity of display, it is important to recognize that the display capabilities of the microcomputer (serving as a delivery system) will necessarily impact the possibilities open to the computerized adaptive test developer. In the development of the CAT-ASVAB battery, considerable effort has been directed toward selecting an appropriate microcomputer (for field testing purposes) to display aptitude test items. We have, however, found that it is sometimes necessary to exclude lengthy paragraph comprehension items from the test and to exclude items written for the technical aptitude tests for which graphic displays are of extreme complexity. However, it is important to note that computerized testing can accommodate new types of tests (e.g. speed and motion) and different item formats (e.g. free response).

With respect to measurement standards, item writing principles that have proven integral to the development of traditional tests must be taken into account in building adaptive tests. It is important to bear in mind that, while all examinees given a paper-and-pencil test receive the same test items and have the opportunity to respond to all items, this is not true for an adaptive test. If a test question is improperly stated, inaccurately keyed, or inappropriate in terms of subject matter for subgroups of examinees, it will be unfairly administered to some, but not all, examinees. For this reason, potential CAT-ASVAB items were carefully screened for adherence to standard item writing rules. In this regard, the role of expert judges has been found to be extremely important.

Recently, two panels of expert judges were charged with screening test items with respect to their conformance to standard item writing principles. One of these two panels was comprised of the same educators and test developers serving to evaluate the test items for appropriate content specification. The other panel was also comprised of twelve members; its participants were professional test developers and minority group educators. The role of this second panel was to ensure that the test was as free as possible of material that could be considered objectionable or unfair to minority groups and to men/women. All potential CAT-ASVAB test items were reviewed

reviewed for content, quality, and sensitivity concerns by the expert judges. Reviewers were instructed to evaluate individual test items for acceptability in terms of adherence to standard item writing principles, such as those described by Wesman (1971), and to indicate their recommendations for overcoming item construction flaws. An account of the item quality and sensitivity reviews involving panels of expert judges will also be included in the CAT-ASVAB Technical Manual.

3. Item Calibration

One of the features of the "main ingredient" for a good CAT is the requirement for well-estimated item parameters for a domain measuring a dominant dimension. In adaptive testing, the ability estimation process and the selection of items for administration depends on estimates of item parameters. These parameters must be determined in advance of adaptive testing. They are typically determined by giving the test items to large samples of examinees, as nearly similar to the test-taking population as possible. According to Green et al., for item response theory calibration via the three-parameter model, a sample of 1,000 cases for each item can be viewed as minimal, and 2,000 cases as adequate. State-of-the-art estimation procedures are often chosen for application via statistical computer programs such as LOGIST.

In the development of CAT-ASVAB, item parameter estimates were obtained for nearly 4,000 items subsequent to pretesting on a recruit sample, and for more than 2,000 items remaining after statistical screening and data collection in a military applicant sample. Item parameters were computed using the estimation program ASCAL (Prestwood, Vale, Massey, and Welsh, 1985). The data collected for more than 2,000 test items on a military applicant sample provided item responses for more than 2,000 examinees per item for calibration purposes. The item parameters determined via this effort will serve as preliminary estimates for use in the field testing of the CAT-ASVAB battery. These item parameter estimates must be considered preliminary, however, until they have been verified for accuracy under computer administration. Consideration of the possible need for recalibrating test items (previously calibrated in paper-and-pencil format) is important in planning the development of a computerized adaptive testing program. A data collection plan is in preparation for accomplishing the verification of item parameters to be used in the CAT-ASVAB battery; the development of new psychometric methods to calibrate future test items is underway.

Another important consideration in the development of computerized adaptive tests stems from an assumption of the IRT models implemented in present adaptive testing methods. The model to be implemented in the CAT-ASVAB battery is the three-parameter model. One of the assumptions of this model is unidimensionality. The outcome of this assumption is that each item included in an adaptive test of a given domain should measure the same unitary construct, in addition to having specific and error variance components associated with it. Green et al. state that while unidimensionality is generally advisable for all tests of ability, it is more critical for adaptive tests. They note that, while the theory is based on unidimensional items, empirical results show suitable results when the items have one dominant dimension.

Measurement of a dominant dimension becomes an important consideration for adaptive testing because different items are administered to examinees. For traditional tests such as the paper-and-pencil ASVAB, all examinees are presented the same test items, and are given an equal opportunity to respond

to them irrespective of their dimensionality. The individualized nature of adaptive tests may raise questions related to appropriate estimation of ability and test fairness.

One of the techniques that has been found to be helpful in studying this problem during the course of CAT-ASVAB development is "full-information factor analysis" (Bock, Gibbons, and Muraki, 1985). This procedure is based on item response theory and is implemented by marginal maximum likelihood estimation. It has been recommended as an exploratory technique in searching for item features that may be responsible for individual differences in cognitive test performance. Irrespective of the procedure used to address the dimensionality issue, this issue should be considered in planning the development of computerized adaptive tests. Addressing the assumption of unidimensionality may entail decisions regarding the elimination of test items that do not measure the dominant dimension, allocation of test items to unidimensional subsets, or the formulation of new psychometric procedures for multidimensional adaptive ability estimation.

Future Directions

As stated earlier, the development of a computerized adaptive testing version of ASVAB involves several technical components. Plans are underway for the collection of additional empirical data to support the development of comparable scores for the computerized and paper-and-pencil versions of the battery, the reliability and validity of the computerized version for its intended purposes, and the verification of test item parameters under computerized administration. The CAT-ASVAB Technical Manual is in preparation, and a CAT-ASVAB delivery system will be field tested this calendar year.

Plans are in progress for the creation of a central test facility to field new CAT-ASVAB test forms, and to analyze and archive CAT-ASVAB test data. Large-scale implementation of this computerized adaptive testing program will permit the investigation of new types of cognitive measures and new item formats. While technical considerations such as those noted above are important to bear in mind in developing a computerized adaptive testing program, the introduction of the microcomputer to the testing process offers a great deal of promise for the expansion of measurement techniques.

REFERENCES

Bock, R.D., Gibbons, R., & Muracki, E. (1985, August). Full-information item factor analysis (MRC Report 85-1). Chicago, ILL: Methodology Research Center/NORC.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.B., & Reckase, M.D. (1984, Winter). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21(4), 347-360.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Prestwood, J.S., Vale, C.D., Massey, R.H., & Welsh, J.R. (1985, September) . Development of an adaptive item pool for the ASVAB (AFHRL-TR-85-19). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Weiss, D.J., & Kingsbury, G.G. (1984, Winter). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21(4), 361-375.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.), Educational measurement (pp.81-129). Washington, DC: American Council on Education.